International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 136-140
Published Online May 2022 in IJEAST (http://www.ijeast.com)

LIVER DISEASE PREDICTION BASED ON GRID SEARCH AND RANDOM FOREST CLASSIFICATION

Saurabh Shinde[1], Kunal Kand[2], Jayesh Sonawane[3], Harshal Jathar[4], Prof. Supriya Bhosale[5]
PCET's Nutan Maharashtra Institute of Engineering and Technology, Pune

***Abstract*—The Medical field is "data rich" and "knowledge poor". This research proposes a Clinical Decision Support System to process this data and early diagnosis of some physiological conditions. With the help of various Machine Learning Techniques we ought to design a CDSS that will assist the doctor to predict disease correctly and thus it may be helpful for patients. This system focuses on to diagnosis of the Liver Diseases. The proposed System uses Decision Tree, Random Forest, Naïve bayes and Support Vector Machine Algorithms for Classification. Finally the proposed system calculates and compares the accuracy of all the four models and demonstrates the best accuracy model for diagnosis of Liver related Diseases. (Abstract)**

***Keywords*-clinical decision support system, patient diagnosis, medical information system, liver disease classification, decision tree optimization, grid search.**

I. INTRODUCTION

Since last two decades, notable advancement has been observed in health monitoring systems inclusive of miscellaneous sectors such as general medicine, clinical decision support system (CDSS), intuitive devices, smart alarm supervising and computer aided diagnostic systems [2].CDSS provides clinicians, patients, or individuals with knowledge and person-specific or population statistics, intelligently strained information which will strengthen health process with better individual patient supervision as well as better inhabitant's health.

According to the annual report published by organization of pharmaceutical producers of India in year 2016 [12], following facts about healthcare came into light,

1) India has 20% of world disease burden. Every 1 in 5 patients worldwide with infectious diseases and Non-communicable disease is an Indian.

2) Only 4.4% of India's gross domestic product is spent on healthcare. 62% of expenses are out-of-pocket. Only 1 in 5 indians is covered by health insurance.

3) India is being called as diabetes capital of the world, there is 123% increase in rate of diabetes which leads to 50% increase in deaths due to diabetes.

Accordingly India needs to take lead and drive patient empowerment system which will accelerates understanding of complex diseases at the molecular level and will provide an integrated environment for analysis of large sets of clinical information and hence forth decreasing India's disease burden.

At national level, there is an increasing trend towards usage of open data from government institutions around the world [3]. The healthcare expenditure can be reduced by using efficient and collaborative health monitoring systems. Thus the old world of limited electronic content with even more limited access to medical data had ownership issues but now new culture which provides ubiquitous content, ready access to critical information has been emerging. This cultural shift now can be used to solve the many ills of the healthcare system.

There is a massive need of early diagnosis systems from patients as well as medical practioner's perspective as early diagnosis provides decision making ability and can potentially be life-saving. The life span of patients suffering from disease might be increased if they are diagnosed in early stages through discoverable symptoms. While considering liver disease, early discovery through symptoms gives various insights like stage of disease, risk factor and damage conditions. Discovery of liver disease is somewhat tricky as partially damaged liver functions nearly like normal functioning.

II. DATASET DESCRIPTION

The dataset from Machine Learning Repository of University of California, Irvine [6] belongs to patients from north east of Andhra Pradesh, India. Total 583 patients were recorded in dataset with liver disease related parameters, 416 patients in dataset are suffering from chronic liver disease and 167 are safe from disease.Data samples are normalized for efficient predictive analysis. Dataset size is quite less which restricts the use of the ensemble methods but random forest algorithm [8] along with the techniques like oversampling, cross-validation and grid search combined with ROC testing makes efficient prediction of disease. In the Table I features or attributes of Indian Liver Patient Dataset (ILPD) [6] has been specified with their respective parametric range of values.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 136-140
Published Online May 2022 in IJEAST (http://www.ijeast.com)

TABLE I
Feature/attribute description of indian
Liver patient dataset (ilpd)

Feature/Attribute	Feature Description	Parameter value	
		Low	High
Gender	Gender of patient	NA	NA
Age	Age of patient	4	90
Total Bilirubin	Bilirubin is a yellow pigment that's found in blood and the stool. Excess bilirubin is a symptom of jaundice.	0.4	75
Direct Bilirubin	Bilirubin is of two types, one that is bound to a certain protein called unconjugated or indirect bilirubin. The other form called direct bilirubin flows directly blood.	0.1	19.7
Alkaline Phosphate	Alkaline phosphatase is an enzyme that's found in the blood and helps in breakdown of proteins. This is an indicator of whether the liver and gall bladder are functioning properly.	63	2110
Alamine Amino Transferase	This enzyme is found in the blood and is a good indicator to verify whether a liver is damaged especially due to cirrhosis and hepatitis.	10	2000
Aspartate Amino Transferase	Low levels of this enzyme are found in the blood. Higher level indicates damage in an organ such as heart or liver.	10	4929
Total Proteins	Total proteins in the body are globulin and albumin. These levels are indicators of liver diseases.	2.7	9.6
Albumin	Albumin is a protein that prevents fluid in blood from leaking out into tissues.	0.9	5.5
Albumin & Globulin Ratio	It's a good indicator of the state of the liver.	0.3	2.8

III. METHODOLOGY

Normalizing and oversampling of dataset will effectively equalize the imbalance between disease-prone and healthy patients. We have implemented our frame work as shown in Figure 1 with Matplotlib for data visualization, Scikit-Learn for algorithms and Pandas for efficient data manipulation. The iPython toolkit [4] allows for a rapid exploration of datasets and algorithms, and is also popular in fields like physics and finance. The Scikit-Learn Python library provides many basic machine-learning capabilities such as clustering, classification and prediction.General framework for prediction uses the flow consisting data preprocessing, sampling/normalization, feature extraction, modelling using classification, and last combined of visualization interpretation, result analysis.

B. Machine Learning algorithms for classification:

1) Decision Tree

In decision tree models [9] the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values are called regression trees.

2) Gaussian Naive Bayes

Naïve Bayes classifier [10] assign class labels to problem instances, represented as vectors of feature values where the class labels are drawn from some finite set.

3) Support Vector Machine

A support vector machine [7] constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space which can be used for classification, regression or other tasks like outliers detection.

4) Random Forest

Randomforest [8] is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

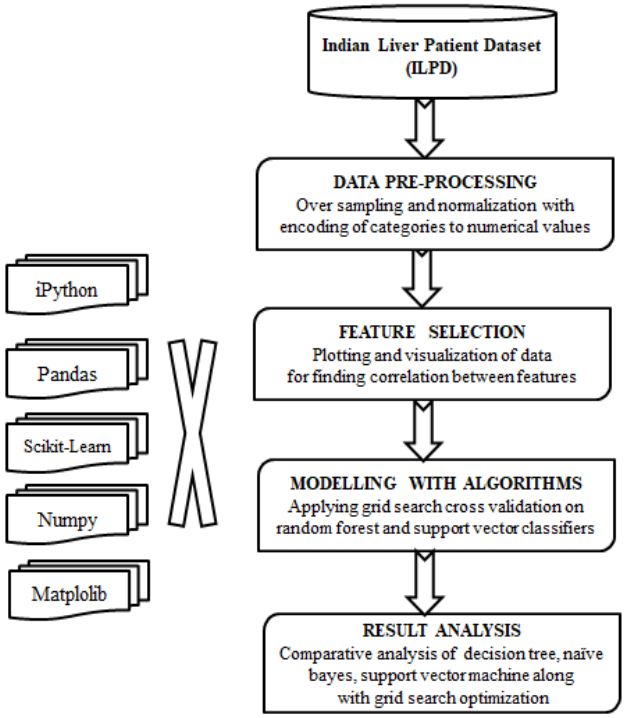

Figure 1.General flow of framework.

C. Clinical data processing with algorithms

1) Data Pre-processing

Present data has variety of parameters with large variation in their ranges thus initially feature scaling is performed. The categorical data is encoded into numerical values for efficient processing also missing values are filled with get_dummies () function from pandas library. Initially the confusion matrix is analyzed for results which gave zero true negative values which means algorithms are always predicting that patient is having a liver disease, then again dataset is tuned further by using oversampling technique in which minority classes are replicated several times so as to account for a difference in the number of healthy livers versus affected livers [1].

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 136-140
Published Online May 2022 in IJEAST (http://www.ijeast.com)

2) **Feature Selection**

To get an insight about correlation between weighted parameter/features, various plots such as join-plot, scatter-plot, bar-plot and heat map are analyzed using Matplotlib library. For example the relation between Total Proteins and Albumin is found by the joint plot between them. Thus related feature can be narrowed down to one single feature. From data visualization using joint plots and scatterplot, we find direct correlation between the following features: first in between Direct Bilirubin and Total Bilirubin, second in between Alamine Amino transferase and Aspartate Amino transferase and third in between Total Proteins and Albumin, fourth in between Albumin and Albumin and Globulin Ratio. A correlation between Total Proteins and Albumin is shown with the Joint-Plot as shown in Figure 2.

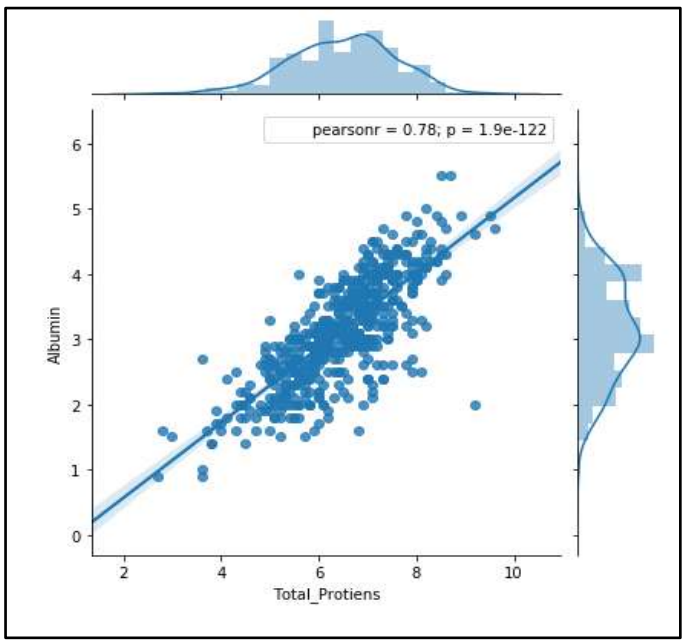

Figure 2.Joint-plot between Total Proteins and Albumin

3) Modelling with algorithms

Total four algorithms are applied on Indian Liver Patients Dataset (ILPD) for comparative study. All algorithms differentiated and cross checked with base majors such as accuracy, recall metric, precision and F-score at initial stage without applying grid-search optimization. random forest and support vector machine techniques performed efficiently with the slightly higher accuracy compared to other algorithms. In lateral stage grid-search optimizer is applied on support vector machine and random forest classifier with initially selected hyper-parameters with scorer object to find optimal parameters. The Random Forests algorithm has several parameters to be adjusted in order to get optimal classifier. Two of those parameters are maximum number of variable used in individual tree and number of trees constructed for classifying new data[4].The result analysis is done using the accuracy measures which are calculated using confusion matrix, means values of true positive, true negative, false positive, false negative samples used for building the accuracy parameters. Accuracy further optimized to maximum value using the **Grid-Search Cross Validation** method along with calculation of ROC-AUC which evaluated combination of parameters that achieved best result.

D. Grid-Search Cross Validation on Classifiers

Different data patterns in machine earning can be generalized using different weights or learning rates and constraints which are called as hyper parameters. For solving the machine learning problem optimally tuning of hyper parameters may aid good results [4]. Optimal model is produced by a selected hyper parameters which minimizes loss function more precisely the difference between true value and estimated value.

Grid-Search is implemented using open source definition in Grid Search CV class present in scikit-learn library [5]. Initially dictionary of hyper parameters is constructed for grid parameter argument. Along with accuracy as a score other scores can also be provided in score argument present in contractor of Grid Search CV. Parallel processing can be done by using multiple cores present which needs no of jobs parameter of Grid Search CV as -1.The Grid Search CV processes will then build and interpret one model for every combination of parameters. Each model is then evaluated using cross validation default value is 3-fold cross validation. Grid-Search outcome can be evaluated using object returned by fit function. The outcome involves best score describing maximum score obtained during optimization and best parameters giving best results. Main characteristics of Grid-Search Cross Validation are listed below:

IV. RESULTS

Initial stage of result analysis yielded different parameters based on which accuracy of algorithms is calculated. Support vector machine and random forest algorithm are more precise and accurate. Support vector machine gives zero true negatives which is incorrect as it is unbalanced and always predicts that the patient is liver disease-prone. Thus grid-Search is used for optimizing and tuning parameters. The accuracy majors of different classification algorithm are evaluated in Table 2consisting of comparison between random forest, decision tree, naive bayes and support vector machine classification techniques with performance evaluation of the same is specified Table 3.In second stage, the performance of different classification techniques on Indian Liver Patient Dataset (ILPD) is evaluated using ROC-AUC testing. Those techniques provide zero value for true negatives which is resolved by grid-search cross validation consisting of tuning of parameters. The AUC for non-grid-search algorithm is 0.4959 which further enhanced to 0.5771 after applying grid-search mechanism on classifier.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 1, ISSN No. 2455-2143, Pages 136-140
Published Online May 2022 in IJEAST (http://www.ijeast.com)

Classification Strategies/Techniques	True Positive	True Negative	False Positive	False Negative	Accuracy
Random Forest	103	21	29	22	71%
Decision Tree	91	33	27	24	64.57%
Naïve Bayes	44	80	2	49	53.14%
Support Vector Machine	124	0	51	0	70%

Table 2.Accuracy majors of different classification techniques on Indian Liver Patient Dataset (ILPD)

Methodology	Precision	Recall	f1-score	Support
Random Forest	70%	71%	70%	175
Decision Tree	66%	65%	65%	175
Naïve Bayes	79%	53%	53%	175
Support Vector Machine	50%	70%	58%	175

Table 3.Performance evaluation different classification techniques on Indian Liver Patient Dataset (ILPD)

In second stage the performance of algorithms is evaluated using ROC-AUC testing. Various algorithms provide zero value for True Negatives which is resolved by Grid-Search cross validation consisting tuning of parameters. The AUC for non-Grid-Searched algorithm was 0.4959 which further increased to 0.5771 after applying Grid-Search on Classifier. Figure 3 below provides difference between ROC curves before and after application of Grid-Search. Formally applying optimization on hyper parameters increases the accuracy score of classifiers modelled using machine learning. This optimization results into best score and best parameters of respective classifier.

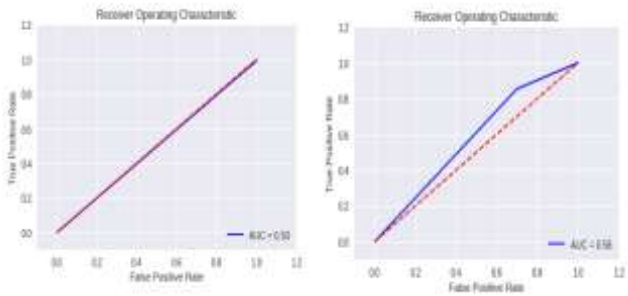

Figure 3.ROC curves before and after Grid Search

For balancing the count of suffering and non-suffering patients of liver disease in database which was leading to more count of False Positives related to wrong prediction, actual dataset was increased using the oversampling method. The results after oversampling studied comparatively between SVM and Random Forest. Random Forest estimated about 0.6954(70%) AUC while for SVM the AUC was around 0.6703(67%). Plot for comparison between two algorithms is given in below figure:

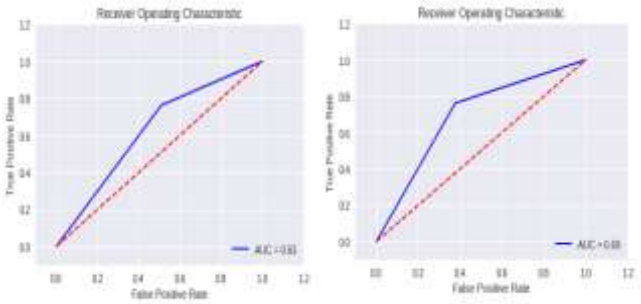

Figure3.ROC curves for SVM and Random Forest algorithms

V. CONCLUSION

Traditional disease diagnostic systems using machine learning for prediction can increase current accuracy using hyper parameter optimization giving the most accurate results. As medical field is digitizing in concern of early diagnosis for using preventive majors there is need of maximizing efficiency and preciseness of existing machine learning system. Thus hyper parmeter optimization techniques like grid search, Random Search, Bayesian optimization, gradient based optimization provide best accuracy score based on best parameters using classifiers. This study used Liver Patients dataset from UCI machine learning repository which was used for modelling of different classifiers, the Random forest algorithm used as best fit with accuracy of 0.72, F1-score of 0.7737, recall metric of 0.7622 and maximum ROC-AUC of about 70%.

VI. REFERENCES

[1]. S. Sontakke, J. Lohokare and R. Dani, "Diagnosis of liver diseases using machine learning," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), Pune, 2017, pp. 129-133

[2]. M. M. Baig, H. G. Hosseini and M. Lindén, "Machine learning-based clinical decision support system for early diagnosis from real-time physiological data," 2016 IEEE Region 10 Conference (TENCON), Singapore, 2016, pp.2943-2946

[3]. H. Ayeldeen, O. Shaker, G. Ayeldeen and K. M. Anwar, "Prediction of liver fibrosis stages by machine learning model: A decision tree approach," 2015 Third World Conference on Complex Systems (WCCS), Marrakech, 2015, pp. 1-6.

[4]. Ramadhan, MuhammaSitanggang, ImasRizky NASUTION, Fahrendi GHIFARI, Abdullah - DEStech Transactions on Computer Science and Engineering, Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency, 10.12783/dtcse/cece2017/14611

[5]. McKinney, Wes. "Data structures for statistical computing in Python." Proceedings of the 9th Python in Science Conference. Vol. 445. 2010.

[6]. UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets

[7]. Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," in IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415-425, Mar 2002.

[8]. J. Ham, Yangchi Chen, M. M. Crawford and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," in IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 492-501, March 2005.

[9]. V. Schetinin et al., "Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications," in IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 3, pp. 312-319, May 2007.

[10]. Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466, Nov. 2006.